The Korean Journal of Applied Statistics (2019) DOI: https://doi.org/10.5351/KJAS.2019.32.4.543
32(4), 543–559

A comparison of imputation methods using nonlinear models

Hyein Kim[a] · Juwon Song[a,1]

[a]Department of Statistics, Korea University

(Received April 16, 2019; Revised May 16, 2019; Accepted May 30, 2019)

Abstract

Data often include missing values due to various reasons. If the missing data mechanism is not MCAR, analysis based on fully observed cases may an estimation cause bias and decrease the precision of the estimate since partially observed cases are excluded. Especially when data include many variables, missing values cause more serious problems. Many imputation techniques are suggested to overcome this difficulty. However, imputation methods using parametric models may not fit well with real data which do not satisfy model assumptions. In this study, we review imputation methods using nonlinear models such as kernel, resampling, and spline methods which are robust on model assumptions. In addition, we suggest utilizing imputation classes to improve imputation accuracy or adding random errors to correctly estimate the variance of the estimates in nonlinear imputation models. Performances of imputation methods using nonlinear models are compared under various simulated data settings. Simulation results indicate that the performances of imputation methods are different as data settings change. However, imputation based on the kernel regression or the penalized spline performs better in most situations. Utilizing imputation classes or adding random errors improves the performance of imputation methods using nonlinear models.

Keywords: missing data, imputation, nonlinear model

1. 서론

대다수의 수집되는 자료에는 여러 가지 원인에 의해 결측이 발생한다. 결측 발생을 방지하고자 연구 계획 및 설계 단계부터 많은 노력을 기울이지만 결측의 문제를 완전히 피하기 어렵다. 결측이 포함된 데이터를 분석할 때, 간단하게 결측값을 제외하고 완전하게 관측된 정보만을 가지고 분석하기 쉬운데 그럴 경우 완전히 관측된 자료가 모집단을 대표한다고 볼 수 없으며 제외된 개체로 인한 정보의 손실로 추정의 정밀도가 약화된다. 또한, 결측이 자료와 상관없이 일어나는 완전임의결측(missing completely at random; MCAR) 메커니즘이 아니라면 결과에 편향이 발생할 수 있다 (Little과 Rubin, 2002). 통상적으로 결측이 하나의 변수에서만 일어나지 않기 때문에 변수가 많은 고차원의 데이터일수록 이 문제는 심화된다. 따라서 결측 자료를 분석하는 여러 방법들이 연구되었으며, 그 중 결측치를 적절한 값으로 채워 넣는 대체 방법은 결측 자료를 완전한 자료로 만들 수 있어 자유롭게 분석할 수 있다는 이점을 지닌다.

[1]Corresponding author: Department of Statistics, Korea University, 145 Anam-ro, Seongbuk-Gu, Seoul 02841, Korea. E-mail: jsong@korea.ac.kr

결측치를 적절한 값으로 채워 넣기 위하여 결측치를 예측하는 통계적인 모델을 세우고 모수를 추정한 후 이를 사용하여 대체를 실시하는 방법들이 제안되어 왔다. 이 때, 사용하는 모델에 따라 평균대체, 회귀대체, 핫덱대체(hotdeck imputation) 등과 같이 부른다. 결측자료 분석의 적절성은 결측자료 메커니즘(missing data mechanism)에 의존하는데 Little과 Rubin (2002)은 결측자료 메커니즘을 완전임의결측(MCAR), 임의결측(missing at random; MAR), 비임의결측(missing not at random; MNAR)으로 분류하였다. 완전임의결측은 결측된 자료와 관측된 자료 모두에 상관없이 결측이 랜덤하게 발생한다는 가정이며 임의결측은 결측이 관측된 자료에는 의존할 수 있으나 결측된 자료에는 상관없이 발생한다는 가정이고 비임의결측은 결측이 발생한 자료 값과 결측 발생이 연관되어 있다는 가정이다. 흔히 사용되는 결측대체 방법들 대부분은 임의결측 가정 하에서 대체를 실시한다

N개의 자료 중 결측이 변수 Y에서만 발생하는 경우를 가정하자. i번째 개체에서 X_i는 완전하게 관측된 변수들의 벡터를 나타내고 관측된 Y_i는 $Y_{\text{obs},i}$로, 결측된 경우 $Y_{\text{mis},i}$로 나타내자. 또한, δ_i는 Y_i의 관측여부를 나타내는데 Y_i가 관측이면 $\delta_i = 1$, Y_i가 결측이면 $\delta_i = 0$으로 표현한다. 즉, 개체 i는 (X_i, Y_i, δ_i)로 표현된다. 관측된 $Y_{\text{obs},i}$에 해당하는 X_i들을 행(row)으로 표현한 행렬을 X_{obs}로, 결측된 $Y_{\text{mis},i}$에 해당하는 X_i들을 행으로 표현한 행렬을 X_{mis}로 나타내자. 본 연구에서는 X와 Y_{obs}값에만 의존하여 결측이 발생하는 임의결측 가정하에서 대체를 실시하는 경우를 고려한다.

먼저 평균대체는 결측값 $Y_{\text{mis},i}$를 다음과 같이 관찰된 Y의 평균값 $\hat{Y}_{\text{mis},i}$으로 대체한다.

$$\hat{Y}_{\text{mis},i} = \#(\text{obs})^{-1} \sum_{j \in \text{obs}} Y_j. \tag{1.1}$$

선형 회귀대체는 반응변수 Y와 설명변수 X들과의 선형 연관 관계를 모델링하여 다음과 같이 대체한다.

$$\hat{Y}_{\text{mis},i} = X_i \hat{\beta}, \tag{1.2}$$

여기서 $\hat{\beta}$은 추정된 Y와 관측된 Y의 오차의 제곱합이 최소가 되게 하는 값으로 $\hat{\beta} = (X_{\text{obs}}^T X_{\text{obs}})^{-1} X_{\text{obs}}^T Y_{\text{obs}}$으로 표현할 수 있다. 이 때, 모수 추정은 Y_i가 관측된 개체들만을 이용하고 정규성과 선형성, 등분산성, 독립성 등 선형 회귀모형 가정을 만족해야 한다. 이 방법은 $Y_{\text{mis},i}$의 예측치로 결측을 대체하는 방법인데 이 예측치에 오차를 더하는 확률적 회귀대체(stochastic regression imputation)로 대체를 실시할 수도 있다. 확률적 회귀대체는 표본 변이를 반영하지 않아 추정치의 분산이 과소 추정되는 식 (1.2)에 의한 대체의 문제를 개선하기 위한 방법으로, 대체된 값에 평균 0, 회귀 잔차의 분산을 가지는 정규분포에서 생성된 오차를 더해주어 결측으로 인한 불확실성을 반영해 결측치를 대체하는 방법을 의미한다 (Little과 Rubin, 2002).

명시적인 모형을 정의하지 않고 적절한 대체값을 찾는 알고리즘 자체에 집중한 방법으로 핫덱(hotdeck)대체나 콜드덱(colddeck)대체 등이 있다. 이 중 흔히 사용되는 핫덱대체는 이미 관측된 개체 중 하나를 뽑아 결측값을 대체하는 방법이다. 가장 기본적인 핫덱대체는 단순임의 핫덱인데 관측된 값들 중 동일한 확률로 한 개를 뽑아 결측값을 대체하는 방법으로 다음과 같이 표현할 수 있다.

$$\hat{Y}_{\text{mis},i} = Y_{\text{obs},j}, \quad \text{with probability } \#(\text{obs})^{-1}. \tag{1.3}$$

이외에도 성별, 지역과 같은 범주형 변수들을 기준으로 비슷한 성향을 가진 개체들로 대체클래스(imputation class)를 형성해 대체클래스 내에서 핫덱 방법을 적용하는 대체클래스를 활용한 핫덱대체 방법이나, 연속형 변수들로 관측값 사이의 거리를 구해 대체하고 싶은 개체와 가장 가까이 있는 관측값으로 대체하는 최근접 이웃 핫덱대체 등이 있다.

선형 회귀대체와 같은 모수적 모형을 이용한 대체 방법들은 비선형의 현실 데이터에 맞지 않을 수 있고 관측된 값 중에서 하나를 랜덤하게 뽑아 대체하는 핫덱대체는 변수간 관계를 적절히 반영해 줄 수 없다

는 한계가 있다. 이런 점들을 개선하기 위해 다양한 비선형 대체 방법들이 연구되었다. Titterington과 Sedransk (1989)는 완전임의 핫덱대체 대신 커널분포를 이용해 핫덱대체에 오차를 더해주는 비모수적 대체 방법인 커널을 활용한 핫덱대체 방법을 제안하였고 Cheng (1994)은 변수간 연관성을 고려하여 예측의 정확성을 높이기 위하여 Nadaraya (1964)의 커널 회귀 추정량을 이용해 비모수적인 대체를 실시하는 커널 회귀대체 방법을 제안하였다. Titterington과 Sedransk (1989)의 커널을 활용한 핫덱대체가 변수들 사이의 연관 관계를 반영하지 못한다는 점과 Cheng (1994)이 제안한 커널 회귀대체가 설명변수와 반응변수간 선형 관계를 가정한다는 문제점을 개선하고자 Aerts 등 (2002)은 완전히 비모수적인 방법인 로컬 리샘플링(local resampling)을 이용한 대체를 제안하였다. 또한, 일부 변수는 선형 관계를 보이고 일부 변수는 비선형 관계를 가지는 경우 모수적인 모델과 비모수적 모델을 결합한 준모수적 방법을 통한 결측대체도 Wang 등 (2004)에 의해 소개되었다. Little과 An (2004)은 결측자료 메커니즘이 임의결측이 되도록 대체 모형을 세우는 경우를 고려하였는데 결측 변수와 나머지 변수들 사이를 모형화할 때 모형의 오지정(misspecification)이 발생할 수 있으므로 이에 덜 영향을 받도록 벌칙 스플라인(penalized spline)을 이용한 결측치 대체를 제안하였다.

다양한 비선형 대체 방법들이 제안되었지만 여러 비선형 대체 방법들 사이의 성능을 비교한 연구는 찾기 힘들었다. 각 논문에서도 대체적으로 기초적인 평균대체나 회귀대체들과 제안한 방법을 비교하고 비선형 대체 방법들 간의 비교가 없었기에 제안된 방법들 간의 성능 비교가 필요하다. 따라서 본 연구에서는 제안된 다양한 비선형 모델을 활용한 대체 방법을 리뷰하고 여러 가지 데이터 설계에서 성능을 비교하였다. 나아가 기존의 Titterington과 Sedransk (1989)를 확장하여 연관된 변수를 2개의 대체클래스로 나눠 각 클래스 안에서 커널을 활용한 핫덱대체를 실시하는 방법을 고려하였다. 또한, Cheng (1994)이 예측평균으로 대체하여 분산을 과소추정하는 한계를 개선하고자 예측평균에 오차를 더하여 대체하는 방법을 제안하였다. 모의실험은 총 5가지의 설계로 이뤄지며 표본 수와 결측의 비율을 달리하여 방법들간의 성능을 비교하였다.

2장에서는 5가지 비선형 모델을 이용한 결측값 대체 방법들의 정의와 의미에 대해 알아보고 3가지 확장된 방법을 제안한다. 소개된 대체 방법들의 성능과 특성을 파악하기 위한 모의실험의 설계 및 결과를 3장에서 소개한다. 마지막으로 4장에서는 연구 결과를 요약하고 추후 연구 방향을 논의한다.

2. 비선형 모델을 이용한 결측값 대체 방법

실제 수집된 데이터가 X와 Y간 선형성을 만족한다는 보장은 할 수 없다. 이에 본 연구에서는 지금까지 소개된 커널, 리샘플링, 준모수 그리고 스플라인 함수를 활용한 비선형 결측값 대체 방법들을 리뷰하고 나아가 기존의 방법들이 가지는 한계를 개선하고자 확장된 방법을 제안한다. 2.1절부터 2.5절까지는 여러 연구에서 제안한 5가지 방법들을 소개하고 2.6절부터 2.8절은 본 연구에서 제안하는 기존 연구를 확장한 대체방법들을 설명한다.

2.1. 커널을 활용한 핫덱대체

Titterington과 Sedransk (1989)는 평균대체(식 (1.1))가 모집단의 평균값 추론에서는 비편향 추정량을 제공할 수도 있지만 평균의 표준오차 추정에서는 편향이 증가함을 주목하였다. 이를 개선하고자 커널분포 추정을 이용해 결측치를 대체하는 핫덱대체 방법을 제안하였다. 전체 개체 N개 중 관측된 개체가 n개인 경우 이 방법은 아래처럼 무작위로 뽑은 관측치에 확률변수 $h * z_i$를 더해 대체하는 방법이다.

$$\hat{Y}_{\text{mis},i} = Y_{\text{obs},j} + hz_i \tag{2.1}$$

이 때 확률변수 z_i는 상호독립적이며, 모두 동일한 확률분포 $K(\cdot)$를 따르며 $\int uK(u)du = 0$이고 $\int u^2K(u)du < \infty$이며 h는 대역폭(bandwidth)을 의미한다. 다시 말해, 이 대체 방법은 커널을 이용해 추정된 모집단의 밀도함수,

$$\hat{f}(Y) = (nh)^{-1} \sum_{j \in \text{obs}} K\left[\frac{Y - Y_j}{h}\right] \tag{2.2}$$

에서 랜덤 추출한 값으로 결측치를 대체하는 것과 같은 의미를 가진다.

2.2. 커널 회귀대체

Cheng (1994)은 Nadaraya (1964)의 커널 회귀 추정량을 활용해 다음과 같이 결측치를 대체하는 방법을 제안하였다.

$$\hat{Y}_{\text{mis},i} = \frac{\sum_{i=1}^{N} K_h(X - X_i)\delta_i Y_i}{\sum_{i=1}^{N} K_h(X - X_i)\delta_i}. \tag{2.3}$$

이 때, $K_h(X) = h^{-1}K((X)/h)$이다.

2.3. 로컬 대체

Titterington과 Sedransk (1989)의 커널을 활용한 핫덱대체는 변수들 간의 연관성을 반영하지 못하며 Cheng (1994)이 제안한 커널 회귀대체는 설명변수와 반응변수간 선형 관계를 가정한다. 변수들간 관계를 반영하면서 완전히 비모수적인 대체 방법으로서 Aerts 등 (2002)은 로컬 리샘플링을 이용한 대체를 제안하였다. 알고리즘은 다음과 같이 로컬 리샘플링과 결측치 대체, 두 단계로 진행한다.

단계 1 관측된 $i = 1, \ldots, n$에 대해서 $\delta_i = 1$일 때, 다음의 분포 $\mathcal{L}(X_i)$에서 $Y^*_{\text{obs},i}$를 추출한다.

$$\mathcal{L}(X_i) = \sum_{j=1}^{N} w_j(X_i) I\{Y_j \le Y_i\}. \tag{2.4}$$

이 때 $w_j(X) = \{\delta_j K_h(X - X_j)\}/\{\sum_{k=1}^{N} \delta_k K_h(X - X_k)\}$이고 $K_h(X) = h^{-1}K((X)/h)$이다.

단계 2 단계 1에서 구한 (X_i, Y_i^*, δ_i)로 식 (2.4)를 이용해 분포 $\mathcal{L}^*(X_i)$를 추정하는데 단계 1과는 달리 관측된 것만이 아닌 전체 표본 $i = 1, \ldots, N$에 대한 분포를 추정한다. $\delta_i = 0$인 $Y_{\text{mis},i}$를 분포 $\mathcal{L}^*(X_i)$에서 추출하여 대체한다.

2.4. 준모수 대체

Wang 등 (2004)은 선형모형과 비모수적인 모형을 결합하여 몇 개의 변수는 선형모형화하고 나머지는 비모수 모형화하여 결측값을 대체하는 준모수적 대체 방법(semiparametric imputation)을 제안하였다. 설명변수 중 선형모델에 사용될 변수는 X로, 비모수적 모델에 사용될 변수는 T로 정의하면 자료는 $(X_i, T_i, Y_i, \delta_i)$, $i = 1, \ldots, N$으로 나타낼 수 있다. 단, X와 T는 완전히 관측된 것으로 가정한다. $\delta_i = 0$인 $Y_{\text{mis},i}$에 대한 대체 방법은 다음과 같다.

$$\hat{Y}_{\text{mis},i} = \mu\left(X_i, \hat{\beta}\right) + \sum_{i=1}^{N} \delta_i W_i(T_i) \left[Y_i - \mu\left(X_i, \hat{\beta}\right)\right] \tag{2.5}$$

여기서 $\hat{\beta}$은 아래의 식을 만족하는 β 값으로 구한다.

$$\min_{\beta} \sum_{i=1}^{n} \delta_i \left\{ \left(Y_i - \sum_{j=1}^{n} W_j(T) Y_j \right) - \left(X_i - \sum_{j=1}^{n} W_j(T) X_j \right)^T \beta \right\}^2$$

이 때, $W(T) = \{K((T-T_j)/h)\}/\{\sum_{j=1}^{N} \delta_j K((T-T_j)/h)\}$이다.

2.5. 벌칙 스플라인을 이용한 결측 대체

Little과 An (2004)은 벌칙 스플라인을 활용한 결측값 대체를 제안하였다. 스플라인 방법은 매듭(knots)을 이용해 유연한 모델링이 가능하지만 매듭을 너무 많이 지정하게 되면 과대적합의 문제가 생기는데, 벌칙 스플라인 (Eilers와 Marx, 1996)은 베이시스 계수의 차분을 이용한 벌칙 함수를 사용함으로써 스플라인 함수의 평활도(smoothness)를 조정해 이 문제를 개선한 방법이다. 기존의 스플라인과 달리 아래의 식과 같이 벌칙(penalty)이 추가된 S를 최소화하는 계수 a_j를 찾아준다.

$$S = \sum_{i=1}^{N} \left(Y_i - \sum_{j=1}^{s} a_j B_j(X_i; q) \right)^2 + \lambda \sum_{j=k+1}^{s} (\Delta^k a_j)^2, \tag{2.6}$$

여기서 $B_j(X_i)$는 주어진 X_i에서 q degree의 j번째 B-spline의 값을 의미한다. N과 s는 각각 관측치의 수와 B-spine의 수이다. λ를 이용해 평활도를 조절하며 벌칙 스플라인을 이용해 분포를 잘 설명하면서도 과대적합이 없는 모델링으로 결측치 대체가 가능하다.

벌칙 스플라인을 이용한 결측 대체에서는 설명변수가 2개, 4개인 경우로 확장하기 위해 일반화 가법 모형(generalized additive model)을 활용한 Marx와 Eilers (1998)의 방법을 적용할 수 있다. Little과 An (2004)은 나아가 공변량의 수가 많아질 때 발생하는 차원의 저주 문제를 개선하기 위해 벌칙 스플라인 성향점수 대체 방법(penalized spline propensity prediction imputation)도 제안하였다. 다차원의 공변량으로 벌칙 스플라인을 적합하는 대신에 일차원의 성향점수를 벌칙 스플라인 모형에 적합하고 나머지 공변량은 선형으로 적합하여 차원의 저주 문제를 해결하였다. 즉,

$$\hat{Y}_{\text{mis},i} = s(P^*) + g(X_{i2}, \ldots, X_{ip}), \tag{2.7}$$

여기서, $P^* = \text{logit}[P(\delta = 1 | X_1, \ldots, X_p)]$이고 $s(P^*)$는 벌칙 스플라인 값이다. 다중공선성 문제를 피하기 위해 $g(\cdot)$에 전체 공변량 중 하나를 제외하며, 현실에서는 P^*를 알 수 없어 로지스틱 회귀를 이용하여 성향점수를 추정해 사용한다.

2.6. 대체클래스별 커널을 활용한 핫덱대체

기존의 Titterington과 Sedransk (1989)의 커널을 활용한 핫덱대체는 $Y_{\text{obs},i}$ 중 하나를 랜덤추출한 뒤 오차를 더해주어 결측값을 대체하기 때문에 Y_{obs}의 범위가 넓을수록 잘못된 값으로 대체될 확률이 높아진다. 따라서 본 연구에서는 X 각 변수를 평균값을 기준으로 2개의 대체클래스로 나눠 각 클래스 내에서 Titterington과 Sedransk (1989)의 방법을 적용하는 방법을 제안한다.

2.7. 확률적 커널 회귀대체

Cheng (1994)의 커널 회귀대체는 커널 회귀선 상의 값으로만 대체되기 때문에 Y의 변동이 과소추정된다. 이런 점을 개선하고자 본 연구에서는 커널 회귀 추정치에 평균 0인 정규분포를 따르는 오차를 더해

주어 결측치를 대체하는 확률적 커널 회귀대체(stochastic kernel regression) 방법을 다음과 같이 제안한다.

$$\hat{Y}_{\text{mis},i} = \frac{\sum_{i=1}^{N} K_h(X - X_i)\delta_i Y_i}{\sum_{i=1}^{N} K_h(X - X_i)\delta_i} + N\left(0, \widehat{\text{Var}}\left(Y - \hat{Y}\right)\right) \tag{2.8}$$

또한, 정규분포가 아닌 실제 잔차를 더해주어 결측치를 대체할 수도 있다. 즉, 관측된 값들과 추정된 값의 차이 중에 랜덤하게 하나를 뽑아 대체된 값에 더하는 방법으로

$$\hat{Y}_{\text{mis},i} = \frac{\sum_{i=1}^{N} K_h(X - X_i)\delta_i Y_i}{\sum_{i=1}^{N} K_h(X - X_i)\delta_i} + \text{deviation}_j \tag{2.9}$$

으로 결측치를 대체한다. 여기서 deviance_j는 관측된 값들의 잔차 중 하나를 의미한다.

2.8. 확률적 벌칙 스플라인 대체

Little과 An (2004)의 벌칙 스플라인을 이용한 결측치 대체 방법도 예측 평균에 의한 대체이므로 Y의 변동이 과소추정될 수 있다. 따라서 식 (2.7)을 이용한 대체에 실제 편차 중 랜덤하게 뽑은 값을 더해주어 결측치를 대체하는 방법을 제안한다. 벌칙 스플라인 대체에 편차를 더해주는 방법으로

$$\hat{Y}_{\text{mis},i} = \sum_{j=1}^{n} a_j B_j(X_i; q) + \text{deviation}_j \tag{2.10}$$

과 벌칙 스플라인 성향점수 대체에 잔차를 더해주는 방법

$$\hat{Y}_{\text{mis},i} = s(P^*) + g(X_2, \ldots, X_p) + \text{deviation}_j \tag{2.11}$$

을 제안한다. 2.7절과 같이 평균 0인 정규분포를 따르는 오차를 더해줄 수도 있지만, 모의실험을 통해 위와 같이 편차를 더해주는 방법이 더욱 적절한 값으로 대체함을 확인하여 정규분포를 따르는 오차를 더해주는 방법은 생략한다.

3. 모의실험

3.1. 모의실험 설계

여러 가지 대체 기법의 성능을 비교하기 위하여 다음과 같은 사항을 고려하여 자료를 생성하였다. 모의실험 1–4의 설계는 Zhang과 Little (2011), 모의실험5는 Kang과 Schafer (2007)의 설계를 참고하였다.

1. 표본의 크기에 따라 성능이 차이가 나는지 살펴보기 위해 표본수는 50, 100, 500으로 설정하였고 각각 1,000번씩 모의실험을 반복하였다.
2. $X = (X_1, \ldots, X_4)$는 서로 독립인 표준정규분포에서 생성하였다.
3. Y는 Table 3.1과 같이 X와의 관계가 선형인 경우(Sim1), 이차항을 가지는 경우(Sim2), 싸인함수(sine function) 형태를 가지는 경우(Sim3), 두 변수의 교호항(Sim4)과 네 개의 변수(Sim5)로 설명되는 평균값을 가지는 정규분포에서 생성하여 총 5가지 상황에서 Y를 생성하였다.
4. 결측의 비율이 낮은 경우와 높은 경우를 가정하기 위하여 결측치의 비율은 10%, 40%인 경우를 고려하였다. 로짓 응답 성향 모형(propensity model)을 이용해 X가 주어졌을 때 관측될 확률을 나타내는 성향점수(propensity score)를 구한 뒤, 성향 점수의 평균보다 작은 값에서 전체 결측치 수의 70%를, 평균보다 큰 값 중에 나머지 30%를 랜덤하게 뽑아 결측치를 생성하였다.

Table 3.1. Data structure

Mean structure	Propensity model
Sim 1: linear mean function $Y\|X \sim N(1+X_1, 1)$	$\text{logit}[P(\delta=1\|X)]$ $=-0.5X_1$
Sim 2: quadratic mean function $Y\|X \sim N(2+2X_1+X_1^2\times M, 1)$	$\text{logit}[P(\delta=1\|X)]$ $=-0.5X_1$
Sim 3: sign mean function $Y\|X \sim N(1+2X_1+2\text{sine}(4X_1)\times M, 1)$	$\text{logit}[P(\delta=1\|X)]$ $=-0.5X_1$
Sim 4: mean function with interaction $Y\|X \sim N(1+X_1+X_2+8X_1X_2\times M, 1)$	$\text{logit}[P(\delta=1\|X)]$ $=0.25X_1-0.5X_2$
Sim 5: mean function with 4 vars. $Y\|X \sim N(210+27.4X_1+13.7X_2+13.7X_3+13.7X_4, 1)$	$\text{logit}[P(\delta=1\|X)]$ $=-X_1+0.5X_2-0.25X_3-0.1X_4$

현실에서는 Y와 X 사이의 실제 구조를 알 수 없기 때문에 제곱항, 교호항, 또는 복잡한 싸인 함수를 포함한 대체모형을 적용하기 힘들다. 따라서 자료의 참모형이 비선형 모형인데 실제로 일차 선형 관계만을 모형에 포함하여 대체를 실시하는 경우를 가정하여 모의실험 2–4에서는 Table 3.1의 식에 $M=0$으로 두어 대체를 실시하였다. 모의실험 5에서는 실제 변수 X가 아닌 아래의 식과 같은 $Z=(Z_1,\ldots,Z_4)$가 관측되었다고 가정하여 대체를 실시하였다.

$$\begin{aligned} Z_1 &= \exp\left(\frac{X_1}{2}\right), \\ Z_2 &= \frac{X_2}{(1+\exp(X_1))+10}, \\ Z_3 &= \left(\frac{X_1X_3}{25}+0.6\right)^3, \\ Z_4 &= (X_2+X_4+20)^2. \end{aligned} \tag{3.1}$$

이는 Z를 관측한 연구자는 정확한 평균 구조나 결측자료 메커니즘을 알지 못해 일반적으로 선형과 로지스틱을 이용해 대체를 실시하는 경우를 반영한다 (Kang과 Schafer, 2007). 따라서 모의실험 1에서만 참 구조 모형으로 대체하고 모의실험 2–5는 대체 모형이 틀린 경우를 가정한다.

모의실험 1–3은 하나의 설명변수를 고려하기 때문에 하나의 설명변수를 선형모형화하면 선형 회귀대체, 비선형 모형화하면 Cheng (1994)의 방법과 동일한 결과를 주는 준모수 대체는 실시하지 않았다. 커널 함수를 활용한 대체에서는 가우시안 커널을 사용하며, 대역폭은 Silverman (1986)방법을 사용하였다. 벌칙 스플라인 대체에서는 3차 스플라인(cubic spline)을 사용하고 설명변수가 1개인 경우 표본 수의 1/3개의 매듭을 고려하고 설명변수가 2개 이상일 때는 4–20개 중 가장 작은 대체값의 평균제곱근오차(root mean square error of individual; RMSE_{ind})를 가지는 매듭의 개수를 찾아 이용하였다.

모의실험 결과는 다음의 기준들을 사용하여 비교하였다.

1. 편향(bias): 참 평균값과 결측 대체 후 추정된 평균값의 차이, $\mu_{\text{true}}-\mu_{\text{estimated}}$.
2. 대체값의 평균제곱근오차:

$$\text{RMSE}_{\text{ind}} = \frac{1}{1000}\sum_{k=1}^{1000}\sqrt{\frac{1}{\#(\text{mis})}\sum_{i\in\text{mis}}\left(Y_{\text{true},i}-\hat{Y}_{\text{impute},i}\right)^2}. \tag{3.2}$$

3. 평균의 평균제곱근오차 증가율(relative root mean square error; RRMSE): 표본 1,000개에서 결측이 발생하지 않았을 때 평균의 평균제곱근오차(RMSE_{bd})보다 결측이 대체되었을 때 평균의 평균제곱근오차(RMSE_{est})의 증가율

$$\text{RRMSE} = 100 \times \frac{\text{RMSE}_{\text{est}} - \text{RMSE}_{\text{bd}}}{\text{RMSE}_{\text{bd}}}. \tag{3.3}$$

4. 신뢰구간 너비 증가율(relative confidence interval widths; RCIW): 결측이 발생하지 않았을 때의 평균의 95% 신뢰구간 너비(CIW_{bd})보다 결측이 대체되었을 때 평균의 95% 신뢰구간 너비(CIW_{est})의 증가율

$$\text{RCIW} = 100 \times \frac{\text{CIW}_{\text{est}} - \text{CIW}_{\text{bd}}}{\text{CIW}_{\text{bd}}}. \tag{3.4}$$

   이 때, 평균 신뢰구간 너비는 1,000개 표본에 대한 $2 * t_{n-1,0.975}\sqrt{\text{Var}(\hat{\mu})}$의 평균을 의미하며 $\text{Var}(\hat{\mu})$는 $\text{Var}(Y)/n$이다.

5. 커버리지 비율(coverage rate; CR): 평균의 95% 신뢰구간 1,000개 중 참값을 포함하는 신뢰구간의 비율.

편향이 0에 가까울수록, 대체값의 평균제곱근오차, 평균의 평균제곱근오차 증가율, 그리고 신뢰구간 너비 증가율이 작을수록 좋은 대체라고 할 수 있다. 커버리지 비율은 반복 개수가 1,000번이므로 0.936보다 작으면 신뢰구간 중 참값을 포함할 확률이 적은 과소포함(undercover) 상태라고 볼 수 있다.

3.2. 모의실험 결과

3.2.1. 모의실험 1 반응변수와 설명변수가 선형적인 관계를 가지는 경우 5가지 비교 기준으로 대체 방법들의 성능을 비교한 결과가 Table 3.2에 나타난다. 결측 비율이 10%일 때는 평균대체와 커널을 활용한 핫덱 대체를 제외한 모든 방법들에서 아주 작은 편향을 보이고, 결측의 비율이 40%로 증가하면 선형 회귀 대체와 벌칙 스플라인에 기반한 대체 방법들의 편향이 작게 나타났다. 대체클래스별 커널을 활용한 핫덱 대체는 기존의 커널을 활용한 핫덱 대체보다 편향이 작아졌으며, 확률적 커널 회귀대체와 확률적 벌칙 스플라인, 확률적 벌칙 스플라인 성향점수 대체 방법은 기존 방법과 유사한 편향을 보인다.

RMSE_{ind}를 살펴보면, 자료의 참 구조와 같은 선형 회귀대체와 벌칙 스플라인 성향점수 대체가 결측의 비율에 상관없이 유사하게 좋은 성능을 보인다. 대체클래스별 커널을 활용한 핫덱 대체는 기존 방법보다 작은 RMSE_{ind}을 가진다. X와의 관계를 직접적으로 반영하지 못하는 평균대체, 커널을 활용한 핫덱 대체는 표본의 수가 많아질 수록 오히려 RRMSE가 커지며, 결측의 비율이 높을 때는 현저하게 성능이 떨어진다.

RCIW은 커널을 활용한 핫덱대체가 거의 모든 경우에서 가장 작았다. 확률적 커널 회귀대체와 확률적 벌칙 스플라인, 확률적 벌칙 스플라인 성향점수 대체는 분산의 과소추정문제를 개선하여 신뢰구간의 감소율이 기존의 대체 방법보다 작았다.

신뢰구간이 과소포함되지 않은 대체 방법들은 진하게 표에 나타냈는데 결측 비율이 낮을 때는 전반적으로 충분한 커버리지 비율을 보이나 결측 비율이 높아지면 평균대체나 커널을 활용한 핫덱 대체는 평균에 대한 신뢰구간의 커버리지 비율이 확연하게 떨어진다. 결측의 비율이 높을 때는 실제 오차가 더해진 확률적 벌칙 스플라인과 확률적 벌칙 스플라인 성향점수 대체가 상대적으로 좋은 커버리지 비율을 보인다.

3.2.2. 모의실험 2 반응변수와 설명변수의 관계가 이차곡선으로 표현되지만 일차식을 가정하여 대

체한 결과가 Table 3.3에 나타난다. 대체클래스별 커널을 활용한 핫덱대체와 벌칙 스플라인에 기반한 방법들이 아주 작은 편향을 가진다. 신뢰구간 너비의 증가율을 보면 거의 모든 경우에서 대체클래스별 커널을 활용한 핫덱대체가 가장 좋은 성능을 보이고 벌칙 스플라인 대체와 벌칙 스플라인 성향점수 대체는 RMSE_{ind}측면에서 가장 좋은 성능을 보인다. 결측의 비율이 10%일 때는 평균대체, 커널을 활용한 핫덱대체와 대체클래스별 커널을 활용한 핫덱대체를 제외한 방법들은 충분한 커버리지 비율을 보인다. 결측의 비율이 40%로 높을 때, 모든 경우 커버리지 비율이 과소포함되었지만 확률적 벌칙 스플라인 대체와 확률적 벌칙 스플라인 성향점수 대체가 가장 높은 커버리지 비율을 보인다.

반응변수가 설명변수의 이차항에 의존하기 때문에 커널 회귀대체나 벌칙 스플라인 대체 방법들의 성능이 좋게 나타났다. 반면에 설명변수와의 관계를 반영하지 못하는 평균대체와 커널을 활용한 핫덱대체는 결측 비율이 높아지면 편향이 증가하여 RRMSE가 매우 커지고 커버리지 비율이 크게 줄어들기 때문에 부적절한 대체 방법으로 나타났다. 확률오차를 포함하도록 확장한 대체 방법은 모의실험 1과 같이 기존의 방법보다 커버리지 비율을 개선함을 알 수 있다.

3.2.3. 모의실험 3 싸인 함수 형태의 평균구조를 바탕으로 데이터를 생성하였으나 선형모형을 가정하여 대체를 실시한 결과는 Table 3.4에 정리하였다. 설명변수와의 관계를 반영하지 못하는 평균대체가 가장 성능이 좋지 않았으며 커널을 활용한 핫덱대체는 커널을 이용한 보정 효과가 약해 성능이 좋지 않으며 이 두 방법은 상대적으로 매우 큰 RRMSE값을 가진다. 벌칙 스플라인에 기반한 방법들이 전반적으로 작은 편향과 높은 커버리지 비율을 보여 좋은 대체법으로 나타났다. 로컬 리샘플링을 활용한 대체는 결측의 비율이 낮을 때는 커널 회귀를 활용한 대체와 비슷한 성능을 보이지만 결측의 비율이 높아지면 선형 회귀대체보다도 더 낮은 커버리지 비율을 가진다.

3.2.4. 모의실험 4 평균구조가 교호항을 포함하는 모의실험 4의 결과가 Table 3.5에 나타난다. 커널 회귀에 기반한 방법들이 RRMSE가 대체적으로 가장 작고 결측의 비율에 상관없이 가장 높은 커버리지 비율을 가진다. 벌칙 스플라인 대체는 다른 모의실험들과 달리 RRMSE 값이 높게 나타나는데 결측 비율이 40%인 경우 평균대체보다도 높은 값을 가진다. 벌칙 스플라인 성향점수 대체는 벌칙 스플라인 대체보다 높은 커버리지 비율을 가지며 결측 비율이 높을 때 벌칙 스플라인이 지나치게 큰 RRMSE를 가지는 점을 개선한다. 준모수 대체방법은 선형 회귀대체보다 전체적으로 성능이 낫지만 커널 회귀대체보다는 좋지 않게 나타나 준모수적 방법의 특징을 잘 나타내고 있다.

3.2.5. 모의실험 5 네 개의 설명변수에 의존하는 평균 구조를 가지나 정확한 변수가 측정되지 않은 경우의 대체 결과는 Table 3.6에 나타난다. 평균대체와 커널을 활용한 핫덱대체, 로컬대체, 그리고 준모수 대체가 큰 편향을 보인다. 벌칙 스플라인 대체는 가장 작은 RMSE_{ind}을 가진다. 결측의 비율이 10%로 낮을 때는 대부분의 대체 방법들이 충분한 커버리지 비율을 보이나 결측 비율이 높아지면 벌칙 스플라인 대체 방법들만 충분한 커버리지 비율을 보인다. 하지만, 표본의 크기가 50이고 결측치 비율이 40%면 관측된 자료의 수는 30개뿐이다. 이 때, 벌칙 스플라인의 매듭이 10개라면 1개의 절편과 9×4개의 계수가 존재해 관측된 값보다 많은 계수를 추정해야 하므로 데이터의 수가 37개보다 적으면 벌칙 스플라인 대체 모형은 모형화가 불가하다는 한계가 존재한다.

4. 결론

수집된 자료의 대다수는 여러가지 이유로 결측이 발생한다. 결측치를 모두 제거하고 분석을 하면 추정치가 편향될 수 있기 때문에 결측값을 적절히 대체하는 게 유용하다. 기존의 모수적 모형을 이용한 대체

Table 3.2. Result of Sim 1

% miss	n	Method	Mean	linear	kernel_hd	kernel_hd_IC	kerne_reg	kernel_reg_S_1	kernel_reg_S_2	local	pspline	pspline_S_2	pspp	pspp_S_2
		Bias	0.054	0.002	0.055	0.004	0.010	0.011	0.010	0.022	0.003	0.004	0.003	0.002
		$RMSE_{ind}$	1.061	0.744	1.588	1.358	0.665	0.645	0.768	1.112	0.692	0.703	0.682	0.768
	50	RRMSE	8.186	3.477	12.869	8.152	3.777	5.475	5.840	7.196	3.603	5.707	3.150	3.405
		RCIW	−5.318	−2.527	0.397	0.471	−3.164	−0.929	−0.907	−0.477	−2.286	−0.020	−2.186	−0.609
		CR	0.928	**0.950**	0.935	**0.939**	**0.947**	**0.949**	**0.945**	**0.943**	**0.948**	**0.947**	**0.950**	**0.954**
		Bias	0.034	−0.002	0.033	0.000	0.002	0.003	0.002	0.009	−0.002	−0.002	−0.002	−0.004
		$RMSE_{ind}$	1.445	1.065	1.968	1.027	1.118	1.482	1.698	1.653	1.065	1.693	1.029	1.083
10%	100	RRMSE	7.151	3.799	13.614	7.143	3.747	6.345	5.093	7.950	3.814	4.974	3.673	4.936
		RCIW	−5.209	−2.514	0.444	0.445	−3.056	−0.767	−0.772	−0.264	−2.360	−0.106	−2.397	−0.563
		CR	0.923	**0.943**	0.914	0.933	**0.942**	**0.939**	**0.939**	**0.938**	**0.943**	**0.940**	**0.943**	**0.938**
		Bias	0.034	−0.002	0.034	−0.003	0.000	0.000	0.000	0.004	−0.002	−0.002	−0.002	−0.001
		$RMSE_{ind}$	1.478	0.878	2.173	1.538	0.907	1.325	1.256	1.466	0.878	1.214	0.883	1.218
	500	RRMSE	16.251	2.472	20.602	7.795	2.307	5.945	3.906	7.084	2.483	4.086	2.362	4.450
		RCIW	−5.149	−2.495	0.278	0.320	−2.789	−0.320	−0.318	−0.238	−2.482	0.006	−2.468	−0.314
		CR	0.882	0.929	0.896	0.922	0.932	0.927	0.934	0.923	0.929	0.930	0.931	0.930
		Bias	0.214	0.003	0.209	0.007	0.036	0.035	0.036	0.082	0.000	0.001	−0.001	0.000
		$RMSE_{ind}$	1.242	0.914	1.867	1.689	0.969	0.999	1.400	2.006	1.751	2.131	0.781	1.172
	50	RRMSE	65.557	17.494	78.638	38.302	20.066	26.825	22.268	43.147	23.131	25.047	18.211	18.790
		RCIW	−24.189	−10.764	0.894	1.443	−13.556	−4.418	−4.191	−2.320	−8.457	0.819	−10.285	−1.283
		CR	0.630	0.877	0.731	0.861	0.855	0.871	0.896	0.827	0.881	0.913	0.875	0.904
		Bias	0.215	−0.001	0.213	0.004	0.026	0.027	0.024	0.062	0.000	−0.002	−0.004	−0.003
		$RMSE_{ind}$	1.435	0.986	2.173	1.410	1.100	1.279	1.274	1.432	1.262	1.465	0.984	1.245
40%	100	RRMSE	100.694	19.510	107.650	37.294	22.613	28.718	26.211	43.868	24.919	28.830	19.918	23.150
		RCIW	−24.227	−10.917	0.259	1.303	−13.277	−3.562	−3.399	−1.753	−9.592	0.038	−10.622	−1.000
		CR	0.481	0.861	0.634	0.864	0.849	0.861	0.881	0.831	0.853	0.894	0.863	0.898
		Bias	0.210	−0.003	0.206	−0.004	0.011	0.011	0.010	0.029	−0.003	−0.004	−0.005	−0.005
		$RMSE_{ind}$	1.510	1.021	2.128	1.673	1.044	1.323	1.507	1.642	1.021	1.490	1.027	1.287
	500	RRMSE	247.577	16.008	247.714	32.122	18.055	24.857	20.674	37.306	16.456	19.389	16.571	19.084
		RCIW	−23.471	−10.516	0.372	1.327	−11.798	−1.480	−1.408	−1.168	−10.361	−0.007	−10.325	−0.172
		CR	0.078	0.862	0.174	0.864	0.858	0.863	0.896	0.828	0.863	0.898	0.864	0.888

mean: 평균대체, linear: 선형 회귀대체, kernel_hd: 커널을 활용한 핫덱대체, kernel_hd_IC: 대체클래스별 커널 핫덱대체, kernel_reg: 커널 회귀대체, kernel_reg_S_1: 정규분포 오차를 더한 확률적 커널 회귀대체, kernel_reg_S_2: 관측된 잔차를 랜덤으로 더한 확률적 커널 회귀대체, local: 로컬 대체, pspline: 벌칙 스플라인을 활용한 대체, pspline_S_2: 관측된 잔차를 랜덤으로 더한 확률적 pspline, pspp: 벌칙 스플라인 성향점수 대체,pspp_S_2 : 관측된 편차를 랜덤으로 더한 확률적 pspp. $RMSE_{ind}$: root mean square error of individual; RRMSE: relative root mean square error; RCIW: relative confidence interval widths; CR: coverage rate.

Table 3.3. Result of Sim 2

% miss	n	Method	Mean	linear	kernel_hd	kernel_hd_IC	kerne_reg	kernel_reg_S_1	kernel_reg_S_2	local	pspline	pspline_S_2	pspp	pspp_S_2
		Bias	0.109	0.011	0.111	0.003	0.034	0.036	0.028	0.059	0.008	0.009	0.006	0.006
		$RMSE_{ind}$	1.312	0.697	1.354	1.561	0.705	0.688	0.915	1.022	0.658	0.851	0.651	0.849
	50	RRMSE	5.274	3.314	9.501	11.405	1.161	1.895	2.156	3.377	1.427	2.219	1.249	1.475
		RCIW	−6.623	−3.455	−1.237	0.623	−2.924	−2.154	−2.057	−1.291	−1.113	−0.431	−0.670	−0.233
		CR	0.898	**0.939**	0.905	0.923	**0.941**	0.935	**0.938**	0.932	**0.946**	**0.947**	**0.949**	**0.951**
		Bias	0.069	−0.001	0.066	0.000	0.016	0.017	0.008	0.031	−0.003	−0.003	−0.003	−0.005
		$RMSE_{ind}$	2.828	2.113	4.140	2.441	1.306	1.666	1.795	2.399	1.128	1.716	1.115	1.121
10%	100	RRMSE	6.407	3.992	13.960	10.917	1.513	2.867	1.756	5.371	1.622	1.782	1.434	1.731
		RCIW	−6.430	−3.406	−1.029	0.036	−2.517	−1.807	−1.743	−1.167	−0.980	−0.372	−0.876	−0.305
		CR	0.906	**0.945**	0.895	0.927	**0.948**	**0.951**	**0.953**	**0.941**	**0.957**	**0.958**	**0.956**	**0.954**
		Bias	0.068	−0.002	0.068	−0.007	0.006	0.006	0.001	0.018	−0.004	−0.005	−0.004	−0.004
		$RMSE_{ind}$	3.423	1.991	4.619	3.296	0.969	1.375	1.296	1.956	0.894	1.233	0.916	1.278
	500	RRMSE	16.556	2.716	21.059	10.428	0.375	1.798	0.575	3.820	0.542	0.745	0.524	0.939
		RCIW	−6.224	−3.275	−1.014	0.291	−1.677	−0.963	−0.894	−0.858	−0.789	−0.086	−0.730	−0.149
		CR	0.865	**0.936**	0.865	0.926	**0.944**	**0.939**	**0.942**	**0.942**	**0.944**	**0.942**	**0.943**	**0.945**
		Bias	0.431	0.071	0.433	0.012	0.140	0.139	0.114	0.228	0.023	0.024	0.022	0.025
		$RMSE_{ind}$	3.013	1.889	3.172	2.864	1.624	1.551	1.720	2.934	1.477	1.888	1.139	1.314
	50	RRMSE	63.848	20.914	73.952	47.381	12.174	15.057	10.935	34.701	8.981	9.384	6.515	7.529
		RCIW	−29.818	−16.432	−8.055	0.356	−13.019	−9.969	−9.632	−6.749	−5.420	−2.734	−5.691	−2.949
		CR	0.545	0.812	0.627	0.833	0.831	0.840	0.868	0.779	0.908	0.919	0.911	0.920
		Bias	0.440	0.068	0.436	0.013	0.107	0.108	0.079	0.189	0.010	0.008	0.008	0.008
		$RMSE_{ind}$	3.589	2.305	4.523	3.888	1.678	1.779	1.795	2.047	1.311	1.492	1.148	1.365
40%	100	RRMSE	105.937	23.836	111.035	50.864	14.979	16.905	13.821	38.295	9.521	10.946	7.404	8.594
		RCIW	−30.440	−17.564	−9.168	−0.600	−11.569	−8.514	−8.163	−6.134	−5.045	−2.314	−4.982	−2.269
		CR	0.392	0.788	0.506	0.799	0.850	0.856	0.877	0.785	0.915	0.928	0.916	0.925
		Bias	0.423	0.046	0.416	−0.004	0.050	0.051	0.029	0.107	−0.004	−0.005	−0.004	−0.004
		$RMSE_{ind}$	2.950	1.576	3.651	2.935	1.055	1.339	1.529	1.899	1.032	1.496	1.041	1.282
	500	RRMSE	268.309	24.168	267.219	42.237	12.066	14.151	7.970	43.743	5.310	5.987	5.350	6.009
		RCIW	−28.669	−16.132	−6.992	1.054	−7.043	−4.075	−3.859	−3.461	−3.433	−0.594	−3.231	−0.483
		CR	0.036	0.782	0.107	0.825	0.879	0.881	0.908	0.781	0.924	0.930	0.925	0.935

mean: 평균대체, linear: 선형 회귀대체, kernel_hd: 커널을 활용한 핫덱대체, kernel_hd_IC: 대체클래스별 커널 핫덱대체, kernel_reg: 커널 회귀대체, kernel_reg_S_1: 정규분포 오차를 더한 확률적 커널 회귀대체, kernel_reg_S_2: 관측된 잔차를 랜덤으로 더한 확률적 커널 회귀대체, local: 로컬 대체, pspline: 벌칙 스플라인을 활용한 대체, pspline_S_2: 관측된 잔차를 랜덤으로 더한 확률적 pspline, pspp: 벌칙 스플라인 성향점수 대체,pspp_S_2 : 관측된 편차를 랜덤으로 더한 확률적 pspp. $RMSE_{ind}$: root mean square error of individual; RRMSE: relative root mean square error; RCIW: relative confidence interval widths; CR: coverage rate.

Table 3.4. Result of Sim 3

% miss	n	Method	Mean	linear	kernel_hd	kernel_hd_IC	kerne_reg	kernel_reg_S_1	kernel_reg_S_2	local	pspline	pspline_S_2	pspp	pspp_S_2
		Bias	0.126	0.023	0.127	−0.002	0.028	0.030	0.030	0.055	0.002	0.003	0.013	0.009
		$RMSE_{ind}$	2.056	1.450	3.800	1.629	0.838	0.705	0.695	2.081	0.645	0.789	1.406	1.100
	50	RRMSE	9.153	2.388	12.528	3.666	2.613	4.107	3.294	5.275	2.501	3.031	1.236	1.918
		RCIW	−5.247	−2.030	0.547	0.447	−2.667	−1.370	−1.375	−0.557	0.225	0.737	−0.937	0.042
		CR	0.924	**0.942**	0.933	**0.941**	**0.943**	**0.939**	**0.949**	**0.939**	**0.950**	**0.953**	**0.949**	**0.946**
		Bias	0.091	0.018	0.088	0.005	0.019	0.021	0.019	0.032	0.002	0.002	0.004	0.004
		$RMSE_{ind}$	2.985	1.662	3.783	1.555	1.421	2.015	2.127	3.048	1.131	1.715	1.183	1.343
10%	100	RRMSE	9.752	2.889	15.017	5.317	2.666	4.216	2.462	6.635	1.172	1.309	1.384	2.200
		RCIW	−5.231	−2.140	0.454	0.224	−2.497	−1.308	−1.268	−0.366	−0.498	0.095	−0.769	0.007
		CR	0.919	**0.937**	0.918	**0.937**	**0.939**	**0.941**	**0.942**	**0.940**	**0.949**	**0.949**	**0.951**	**0.946**
		Bias	0.090	0.018	0.091	−0.002	0.011	0.011	0.010	0.022	0.000	0.000	0.000	0.002
		$RMSE_{ind}$	3.045	1.782	4.372	2.632	1.142	1.571	1.627	2.254	0.960	1.360	1.140	1.599
	500	RRMSE	28.828	3.369	33.668	4.793	1.754	3.048	1.025	7.872	0.364	0.417	0.201	0.497
		RCIW	−5.206	−2.176	0.268	0.221	−1.969	−0.987	−0.967	−0.300	−0.786	−0.113	−0.715	0.005
		CR	0.838	**0.938**	0.847	**0.941**	**0.940**	**0.944**	**0.951**	0.930	**0.949**	**0.950**	**0.953**	**0.953**
		Bias	0.521	0.103	0.520	0.006	0.138	0.138	0.129	0.2 47	0.011	0.013	0.040	0.030
		$RMSE_{ind}$	2.744	1.942	3.884	2.240	1.430	1.676	1.959	4 .104	4.198	4.367	1.657	2.094
	50	RRMSE	88.840	18.889	101.398	26.011	22.596	26.282	21.776	4 8.526	70.799	71.289	10.075	11.641
		RCIW	−24.807	−9.752	0.134	0.548	−12.138	−6.677	−6.467	−3.099	8.567	10.773	−4.331	0.588
		CR	0.523	0.856	0.656	0.880	0.840	0.855	0.870	0.785	0.905	0.915	0.913	0.920
		Bias	0.542	0.113	0.539	0.013	0.125	0.126	0.112	0.212	−0.001	−0.003	0.011	0.005
		$RMSE_{ind}$	2.988	1.704	4.525	2.502	1.357	1.835	1.843	3.159	2.346	2.612	1.722	2.321
40%	100	RRMSE	144.861	26.268	150.519	23.862	28.019	31.823	26.016	57.876	15.647	16.934	10.626	10.868
		RCIW	−24.882	−9.983	−0.495	0.610	−11.124	−5.958	−5.808	−2.104	0.062	2.390	−2.634	0.979
		CR	0.308	0.843	0.473	0.895	0.835	0.834	0.864	0.775	0.918	0.920	0.925	0.930
		Bias	0.538	0.117	0.532	−0.003	0.079	0.079	0.066	0.147	0.001	0.000	0.000	0.001
		$RMSE_{ind}$	2.982	1.743	3.998	2.312	1.236	1.538	1.728	2.096	1.057	1.521	1.052	1.318
	500	RRMSE	374.472	54.691	372.702	22.964	31.827	34.281	24.494	80.501	6.395	6.697	5.598	6.503
		RCIW	−24.201	−9.734	−0.588	0.776	−8.352	−4.257	−4.097	−1.417	−3.209	−0.471	−2.912	−0.147
		CR	0.011	0.717	0.035	0.893	0.824	0.823	0.866	0.675	**0.937**	**0.938**	**0.937**	**0.943**

mean: 평균대체, linear: 선형 회귀대체, kernel_hd: 커널을 활용한 핫덱대체, kernel_hd_IC: 대체클래스별 커널 핫덱대체, kernel_reg: 커널 회귀대체, kernel_reg_S_1: 정규분포 오차를 더한 확률적 커널 회귀대체, kernel_reg_S_2: 관측된 잔차를 랜덤으로 더한 확률적 커널 회귀대체, local: 로컬 대체, pspline: 벌칙 스플라인을 활용한 대체, pspline_S_2: 관측된 잔차를 랜덤으로 더한 확률적 pspline, pspp: 벌칙 스플라인 성향점수 대체,pspp_S_2 : 관측된 편차를 랜덤으로 더한 확률적 pspp. $RMSE_{ind}$: root mean square error of individual; RRMSE: relative root mean square error; RCIW: relative confidence interval widths; CR: coverage rate.

Table 3.5. Result of Sim 4

% miss	n	Method	Mean	linear	kernel_hd	kernel_hd_IC	kerne_reg	kernel_reg_S_1	kernel_reg_S_2	local	semi	pspline	pspline_S_2	pspp	pspp_S_2
		Bias	−0.027	−0.049	−0.048	−0.014	−0.029	−0.029	−0.031	0.011	−0.042	−0.051	−0.057	−0.043	−0.018
		$RMSE_{ind}$	10.348	12.740	12.575	12.660	5.326	5.085	5.311	10.597	10.594	9.820	8.581	13.273	15.333
	50	RRMSE	6.632	7.352	10.297	10.989	0.613	0.757	0.507	5.234	4.815	7.079	8.136	8.904	7.155
		RCIW	−5.814	−5.057	−0.491	−0.768	−2.816	−2.748	−2.750	−2.696	−4.878	−3.855	−2.267	−1.669	−0.291
		CR	0.916	0.915	0.921	0.927	**0.940**	**0.940**	**0.939**	0.930	0.918	0.916	0.918	0.930	**0.938**
		Bias	0.027	0.012	0.033	0.032	0.000	0.000	0.000	0.091	0.020	0.013	0.008	0.011	0.018
		$RMSE_{ind}$	9.067	8.093	13.114	11.686	2.108	2.057	2.070	11.026	7.902	8.093	8.418	6.089	6.060
10%	100	RRMSE	5.815	5.824	12.873	8.667	0.522	0.668	0.738	5.866	4.819	5.684	7.225	3.517	3.381
		RCIW	−5.831	−5.341	−0.497	−0.642	−2.042	−1.975	−1.990	−2.187	−5.075	−4.220	−1.990	−3.059	−1.095
		CR	0.923	0.929	0.923	0.931	**0.947**	**0.945**	**0.946**	0.933	0.933	0.927	**0.936**	**0.936**	**0.939**
		Bias	0.020	0.000	0.021	0.023	0.005	0.006	0.006	0.035	0.003	−0.001	0.005	0.011	0.046
		$RMSE_{ind}$	7.203	6.565	10.484	10.908	1.316	1.745	1.535	9.085	6.683	6.565	9.499	6.637	9.504
	500	RRMSE	4.799	4.605	9.777	9.078	0.122	0.052	0.157	5.733	4.428	5.651	9.679	4.308	6.179
		RCIW	−5.457	−5.253	−0.267	−0.260	−0.951	−0.900	−0.896	−1.391	−5.081	−4.436	−1.123	−3.302	−0.816
		CR	0.913	0.917	0.918	0.917	**0.937**	**0.937**	**0.938**	0.921	0.918	0.911	0.910	0.921	0.922
		Bias	0.015	−0.194	0.014	−0.016	−0.019	−0.021	−0.020	0.291	−0.009	−0.211	−0.223	−0.047	−0.007
		$RMSE_{ind}$	6.850	9.464	8.661	12.785	2.847	3.034	2.942	8.533	7.388	8.435	9.393	4.621	4.119
	50	RRMSE	28.247	35.754	49.012	40.236	5.102	5.325	5.267	29.166	20.884	50.279	53.224	32.198	31.344
		RCIW	−25.941	−20.786	−3.227	−4.113	−12.428	−12.136	−12.131	−11.31	−21.184	−13.284	−4.76	−14.534	−5.603
		CR	0.728	0.714	0.813	0.830	0.888	0.891	0.891	0.826	0.792	0.721	0.759	0.799	0.852
		Bias	0.084	−0.141	0.090	0.082	0.004	0.004	0.004	0.383	0.022	−0.145	−0.140	0.032	0.082
		$RMSE_{ind}$	8.021	8.680	10.736	12.370	4.294	4.163	4.275	10.882	8.429	8.680	10.470	7.639	11.310
40%	100	RRMSE	30.084	39.677	43.118	45.541	5.817	6.081	5.913	35.973	25.532	47.506	50.112	33.496	33.028
		RCIW	−24.885	−21.818	−3.001	−2.518	−9.368	−9.138	−9.171	−8.851	−20.965	−15.834	−3.417	−15.378	−4.680
		CR	0.743	0.736	0.816	0.812	0.905	0.907	0.910	0.812	0.786	0.745	0.803	0.793	0.854
		Bias	0.098	−0.152	0.105	0.088	0.008	0.008	0.008	0.197	−0.007	−0.173	−0.178	−0.005	0.040
		$RMSE_{ind}$	7.177	7.116	10.666	9.608	1.876	2.045	2.000	9.813	7.653	7.231	9.846	6.201	8.129
	500	RRMSE	24.605	33.859	38.839	36.371	−0.005	0.244	−0.192	37.083	17.254	43.584	48.897	20.799	23.666
		RCIW	−23.844	−22.746	−1.274	−1.489	−4.562	−4.373	−4.377	−5.354	−21.973	−16.152	−0.030	−15.446	−3.001
		CR	0.740	0.725	0.822	0.819	0.928	0.930	0.926	0.812	0.791	0.733	0.797	0.811	0.857

mean: 평균대체, linear: 선형 회귀대체, kernel_hd: 커널을 활용한 핫덱대체, kernel_hd_IC: 대체클래스별 커널 핫덱대체, kernel_reg: 커널 회귀대체, kernel_reg_S_1: 정규분포 오차를 더한 확률적 커널 회귀대체, kernel_reg_S_2: 관측된 잔차를 랜덤으로 더한 확률적 커널 회귀대체, local: 로컬 대체, semi: 준모수 대체, pspline: 벌칙 스플라인을 활용한 대체, pspline_S_2: 관측된 잔차를 랜덤으로 더한 확률적 pspline, pspp: 벌칙 스플라인 성향점수 대체,pspp_S_2 : 관측된 편차를 랜덤으로 더한 확률적 pspp. $RMSE_{ind}$: root mean square error of individual; RRMSE: relative root mean square error; RCIW: relative confidence interval widths; CR: coverage rate.

Table 3.6. Result of Sim 5

% miss	n	Method	Mean	linear	kernel_hd	kernel_hd_IC	kerne_reg	kernel_reg_S_1	kernel_reg_S_2	local	semi	pspline	pspline_S_2	pspp	pspp_S_2
		Bias	1.155	0.293	1.122	0.279	0.320	0.319	0.279	1.419	2.267	−0.074	−0.063	0.106	−0.020
		$RMSE_{ind}$	29.501	24.503	43.336	30.100	14.389	20.383	9.786	41.958	41.930	5.607	6.913	21.565	19.666
	50	RRMSE	8.606	4.096	13.173	9.464	2.298	2.419	2.868	11.730	15.778	0.484	0.491	2.243	2.045
		RCIW	−5.304	−0.820	0.328	0.461	−2.345	−1.892	−1.840	−0.804	0.471	0.262	0.239	1.279	1.957
		CR	0.918	**0.948**	0.922	0.934	**0.941**	**0.947**	**0.944**	0.929	0.924	**0.956**	**0.953**	**0.952**	**0.957**
		Bias	0.811	0.167	0.762	0.130	0.069	0.069	0.069	1.123	1.024	−0.031	−0.030	0.095	0.004
		$RMSE_{ind}$	24.206	22.068	52.186	42.609	10.291	12.359	13.072	60.228	8.080	4.895	4.241	17.027	17.168
10%	100	RRMSE	8.227	2.974	13.841	8.815	1.118	1.393	1.016	10.990	7.254	0.041	0.023	1.721	1.792
		RCIW	−5.159	−1.078	0.525	0.635	−1.659	−1.451	−1.465	−0.644	−0.760	0.085	0.175	0.734	1.130
		CR	0.922	**0.947**	0.929	**0.936**	**0.951**	**0.950**	**0.951**	0.931	**0.938**	**0.960**	**0.959**	**0.955**	**0.955**
		Bias	0.729	0.194	0.736	0.096	−0.028	−0.029	−0.019	0.913	0.230	−0.041	−0.044	0.068	0.026
		$RMSE_{ind}$	34.538	20.761	45.535	42.339	9.885	10.615	10.113	49.989	11.796	7.235	10.034	14.228	16.811
	500	RRMSE	14.101	2.151	18.767	7.028	−0.305	−0.275	−0.278	21.922	1.517	0.003	0.107	4.032	4.635
		RCIW	−5.170	−1.348	0.293	0.392	−1.001	−0.957	−0.953	−0.963	−0.883	0.008	0.110	1.410	1.694
		CR	0.910	**0.953**	0.916	**0.950**	**0.959**	**0.959**	**0.959**	0.901	**0.951**	**0.959**	**0.958**	**0.960**	**0.961**
		B ias	4.493	0.572	4.624	0.525	1.541	1.693	1.573	5.774	12.719	−0.349	−0.359	0.555	0.464
		$RMSE_{ind}$	34.112	24.029	35.928	44.533	18.935	19.483	31.008	47.312	84.315	9.589	9.676	17.081	17.743
	50	RRMSE	58.282	18.811	74.146	36.948	19.636	23.865	20.715	79.176	212.9	4.279	4.458	9.760	9.545
		RCIW	−24.123	−2.386	0.831	1.779	−11.907	−8.820	−8.766	−4.826	28.078	2.153	2.429	1.968	2.800
		CR	0.667	0.893	0.752	0.855	0.851	0.866	0.864	0.698	0.546	**0.952**	**0.952**	**0.946**	**0.954**
		Bias	4.707	0.830	4.770	0.845	0.541	0.618	0.621	6.084	7.676	−0.213	−0.222	0.467	0.369
		$RMSE_{ind}$	34.785	28.881	55.494	42.169	14.516	16.432	16.379	49.832	38.229	7.329	9.144	10.715	12.676
40%	100	RRMSE	85.351	20.049	97.695	36.850	9.162	11.008	11.597	120.49	177.446	3.887	3.865	5.407	5.564
		RCIW	−23.559	−2.648	0.986	2.487	−7.862	−6.815	−6.713	−3.434	14.440	1.244	1.593	0.769	1.865
		CR	0.560	0.893	0.671	0.875	0.901	0.902	0.894	0.551	0.574	**0.951**	**0.953**	**0.945**	**0.943**
		Bias	4.619	1.152	4.574	0.841	−0.023	−0.018	0.012	5.439	3.285	−0.145	−0.148	0.152	0.132
		$RMSE_{ind}$	40.101	23.180	51.825	46.368	3.780	13.834	14.828	45.906	32.222	11.336	12.468	6.221	7.443
	500	RRMSE	220.422	37.935	224.708	44.879	2.500	2.716	2.657	275.105	153.9	5.912	6.395	4.812	4.888
		RCIW	−23.186	−3.999	0.667	1.868	−4.293	−4.021	−4.047	−4.131	3.295	1.234	1.688	1.058	1.772
		CR	0.137	0.835	0.282	0.846	**0.945**	**0.943**	**0.946**	0.149	0.526	**0.947**	**0.947**	**0.952**	**0.952**

mean: 평균대체, linear: 선형 회귀대체, kernel_hd: 커널을 활용한 핫덱대체, kernel_hd_IC: 대체클래스별 커널 핫덱대체, kernel_reg: 커널 회귀대체, kernel_reg_S_1: 정규분포 오차를 더한 확률적 커널 회귀대체, kernel_reg_S_2: 관측된 잔차를 랜덤으로 더한 확률적 커널 회귀대체, local: 로컬 대체, semi: 준모수 대체, pspline: 벌칙 스플라인을 활용한 대체, pspline_S_2: 관측된 잔차를 랜덤으로 더한 확률적 pspline, pspp: 벌칙 스플라인 성향점수 대체,pspp_S_2 : 관측된 편차를 랜덤으로 더한 확률적 pspp. $RMSE_{ind}$: root mean square error of individual; RRMSE: relative root mean square error; RCIW: relative confidence interval widths; CR: coverage rate.

방법들은 비선형성을 가지는 현실 데이터에 적용하기엔 한계가 있으므로 비선형 모형을 사용한 대체 방법을 사용하는 게 바람직할 것이다. 따라서 본 논문에서는 기존에 연구되었던 다양한 비선형 대체 방법들을 여러 모의실험 설계에 적용해 성능을 비교하였다. 나아가 기존 방법의 편향을 줄이기 위해 대체군을 활용하거나 분산의 과소추정 문제를 개선하기 위해 랜덤 오차를 더하는 확률적인 대체를 제안하였다.

모의실험에 따르면 평균대체나 선형 회귀대체는 참 평균 구조가 선형일 때는 타 방법들과 비슷한 성능을 보이지만, 이차항과 싸인 함수에 의존하거나 결측 비율이 높아지면 성능이 현저히 떨어진다. 커널을 활용한 핫덱대체는 결측 비율이 커지면 RRMSE가 과도하게 증가하였지만 전반적으로 다른 방법들보다 참 신뢰구간 길이와 차이가 적게 나타났다. 커널 회귀대체는 설명변수가 여러 개인 모의실험 5를 제외하고는 전반적으로 성능이 좋았으며 특히 교호항이 연관되어 있는 경우에는 다른 대체 방법들보다 뛰어난 성능을 보였다. 로컬 리샘플링 대체는 평균 구조가 교호항에 의존하는 모의실험 4를 제외하고는 선형 회귀대체보다 낮은 커버리지 비율을 보인다. 준모수적 대체는 설명변수가 많은 모의실험 5에서는 좋은 성능을 보여주지 못했지만 설명변수가 2개인 모의실험 4에서는 선형 회귀대체와 커널 회귀대체의 중간 성능을 보였다. 교호항이 있는 경우를 제외하고는 벌칙 스플라인 대체와 벌칙 스플라인 성향점수 대체는 결측 비율이 증가하여도 커버리지 비율의 감소폭이 가장 작았으며 전반적으로 좋은 성능을 보였다.

변수들 간에 선형적인 연관성을 지닌 자료에서는 비선형 방법들 뿐 아니라 선형 회귀대체도 좋은 성능을 보여 강건한(robust) 대체를 제공하였다. 따라서 만약 데이터의 구조가 선형적이라면 간단한 선형 회귀대체를 이용하는 것이 가장 효율적일 것이다. 하지만 결측의 비율이 높을 때는 관측된 편차를 더해주는 확률적 커널 회귀대체나 확률적 벌칙 스플라인 대체, 확률적 벌칙 스플라인 성향점수 대체를 사용하는 것도 대안이 될 것이다.

변수들 간에 선형이 아닌 복잡한 관계를 가지는 경우에는 벌칙 스플라인 대체나 벌칙 스플라인 성향점수 대체가 평균을 참값에 가까이 추정함을 알 수 있다. 만약 두 변수의 교호항과 연관된 평균 구조를 가진 자료라고 생각되면 벌칙 스플라인을 활용한 방법보다 커널 회귀대체와 확률적 커널 회귀대체가 더 적절한 대체를 할 것으로 기대된다. 한편, 관측자가 변수들 간의 어떤 구조도 알지 못한다고 하면 벌칙 스플라인 대체를 사용하는 것을 추천한다. 하지만 이 방법은 적절한 매듭의 개수를 찾아주는 과정이 필요하기 때문에 결측의 비율이 크지 않다면 선형 회귀나 커널 회귀에 기반한 대체 방법도 간단하게 적용해 볼 수 있겠다.

벌칙 스플라인의 장점은 비교적 충분한 수의 매듭을 지정해도 벌점 함수가 과적합의 문제를 막는다고 알려져 있다. 하지만 설명변수의 수가 많을 때 매듭을 10개만 해도 RRMSE가 600을 넘기며 과적합의 문제를 피할 수 없었다. 따라서 본 연구에서는 설명변수가 2개 이상인 경우, $\mathrm{RMSE}_{\mathrm{ind}}$을 기준으로 적절한 매듭의 수를 찾아주었다. 표본 수(50, 100, 500)별로 최적 매듭 개수의 평균은 5, 6, 11개였다.

커널을 활용한 핫덱대체를 확장한 대체클래스별 커널을 활용한 핫덱대체는 기존의 방법보다 성능이 월등히 개선되었으며 커널 회귀대체를 확장한 확률적 커널 회귀대체는 평균값의 분산을 과소추정하는 문제를 개선하였다. 확률적 벌칙 스플라인과 확률적 벌칙 스플라인 성향점수 대체는 평균값의 분산을 과소 추정하는 정도를 줄여 거의 모든 경우에서 기존 방법보다 커버리지 비율이 증가했다.

본 연구에서는 결측 비율과 평균구조를 고려하여 모의실험을 시행하였지만, 결측자료 메커니즘에 연관된 관측 확률인 성향점수의 구조에 따른 성능의 변화를 살펴보지는 못했다. 비선형 방법 중 벌칙 스플라인 성향점수 대체는 평균 구조나 성향점수 모형 둘 중 하나만 참 구조와 일치하면 모 평균으로 수렴하는 이중 강건한(doubly robust) 성질을 가진다. 따라서 이후의 연구에서는 성향점수의 구조를 다양하게 설계하여 모의실험 5와 같이 평균구조와 성향점수 모두 참 구조와 일치하지 않을 때 비선형 방법들과 성능을 더 자세히 비교할 수 있을 것으로 기대된다.

References

Aerts, M., Claeskens, G., Hens, N., and Molenberghs, G. (2002). Local multiple imputation, *Biometrika*, **89**, 375–388.

Cheng, P. E. (1994). Nonparametric estimation of mean functionals with data missing at random, *Journal of the American Statistical Association*, **89**, 81–87.

Eilers, P. H. C. and Marx, B. D. (1996). Flexible smoothing with B-splines and penalties, *Statistical Science*, **11**, 89–121.

Kang, D. Y. J. and Schafer, J. L. (2007). Demystifying double robustness: a comparison of alternative strategies for estimating a population mean from incomplete data, *Statistical Science*, **22**, 523–539.

Little, R. J. A. and An, H. (2004). Robust likelihood-based analysis of multivariate data with missing values, *Statistica Sinica*, **14**, 949–968.

Little, R. J. A. and Rubin, D. B. (2002). *Statistical Analysis with Missing Data* (2nd ed.), Wiley, New York.

Nadaraya, E. A. (1964). On estimating regression, *Theory of Probability and its Application*, **9**, 141–142.

Marx, B. D. and Eilers, P. H. C. (1998). Direct generalized additive modeling with penalized likelihood, *Computational Statistics & Data Analysis*, **28**, 193–209.

Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*, Chapman and Hall, London.

Titterington, D. M. and Sedransk, J. (1989). Imputation of missing values using density estimation, *Statistics & Probability Letters*, **8**, 411–418.

Wang, Q., Linton, O., and Hardle, W. (2004). Semiparametric regression analysis with missing response at random, *Journal of the American Statistical Association*, **99**, 334–345.

Zhang, G. and Little, R. J. A. (2011). A comparative study of doubly robust estimators of the mean with missing data, *Journal of Statistical Computation and Simulation*, **81**, 2039–2058.

비선형 모델을 이용한 결측 대체 방법 비교

김혜인[a] · 송주원[a,1]

[a]고려대학교 통계학과

(2019년 4월 16일 접수, 2019년 5월 16일 수정, 2019년 5월 30일 채택)

요 약

자료에는 다양한 원인에 의해 결측이 발생한다. 만약 결측치를 제외하고 완전히 관찰된 자료만으로 분석을 실시한다면 결측자료 메커니즘이 완전임의결측이 아닌 경우 결과에 편향이 발생하거나 제외된 개체로 인한 정보의 손실로 추정의 정밀도가 약화된다. 결측이 하나의 변수에서만 일어나지 않기 때문에, 자료에 변수가 많을 수록 이 문제는 심화된다. 문제를 개선하기 위해 결측치를 대체하는 여러가지 방법들이 제안되었다. 하지만 모수적인 모형을 이용한 대체 방법들은 가정에 위배되는 현실 데이터에는 적합하지 않다. 따라서 본 연구에서는 자료의 분포 가정에 덜 영향을 받는 커널, 리샘플링, 스플라인 방법을 활용한 비선형 대체 방법들을 리뷰하고 필요한 경우 기존의 비선형 대체 방법에 대체클래스를 사용하여 대체값의 정확도를 높이거나 랜덤성을 가지는 오차를 더해주어 추정치의 분산이 적게 추정되는 문제를 개선하는 확장된 결측 대체 방법을 제안한다. 본 연구에서 고려한 여러 가지 대체 방법들은 다양한 모의자료 설계 하에서 성능을 비교하였다. 모의실험 결과, 비선형 대체 방법들은 각 설계 하에 다른 성능을 보이며 전반적으로 커널 회귀나 스플라인을 활용한 대체 방법들이 좋은 성능을 보였다. 더불어, 확장된 대체 방법은 기존의 대체 방법이 가지는 문제점을 개선함을 확인할 수 있었다.

주요용어: 결측자료, 대체, 비선형모형

[1]교신저자: (02841) 서울특별시 성북구 안암로 145, 고려대학교 통계학과. E-mail: jsong@korea.ac.kr